Sự tương đồng và khoảng cách (tiếp theo)

Độ đo tương đồng theo thời gian

- Dữ liệu thời gian bao gồm
  - Một thuộc tính ngữ cảnh biểu diễn thời gian.
  - Một hoặc nhiều thuộc tính hành vi biểu diễn các đặc tính thay đổi theo một khoảng thời gian.
- Dữ liệu thời gian có thể được biểu diễn bằng một chuỗi thời gian liên tục hoặc một dãy rời rạc, tùy thuộc vào ứng dụng.
- Cách biểu diễn theo dãy rời rạc có thể xem là dạng rời rạc của cách biểu diễn theo chuỗi thời gian liên tục.

Độ đo tương đồng chuỗi thời gian

- Việc thiết kế độ đo tương đồng chuỗi thời gian rất phụ thuộc vào từng ứng dụng. Một số các yếu tố ảnh hưởng như sau.
  - Scaling và chuyển vị thuộc tính hành vi.
  - Chuyển vị thuộc tính thời gian (ngữ cảnh).
  - Scaling thuộc tính thời gian (ngữ cảnh).
  - Sự không liên tục khi so khớp.

Độ đo tương đồng chuỗi thời gian

- Ảnh hưởng của việc chuẩn hóa thuộc tính hành vi.
- Các vấn đề về scaling và chuyển vị với các thuộc tính hành vi thường dễ xử lý hơn các thuộc tính ngữ cảnh với các phép chuẩn hóa sau.
  - Chuyển vị thuộc tính hành vi.
  - Scaling thuộc tính hành vi.

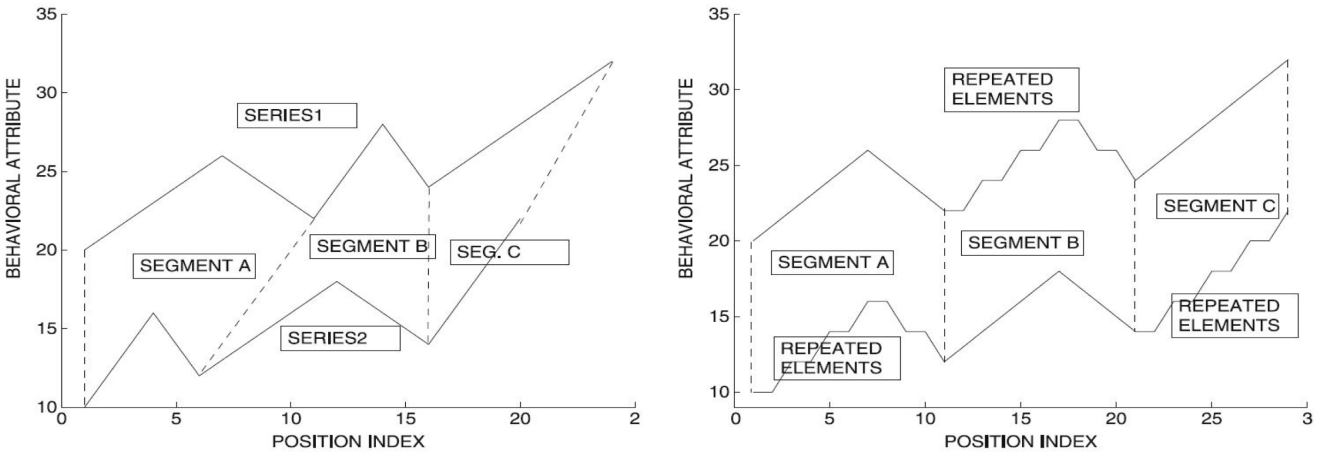
Chuẩn Lp

- Chúng ta có thể xem mỗi chuỗi thời gian là dạng dữ liệu đa chiều với mỗi mốc thời gian là một chiều và từ đó áp dụng chuẩn Lp.

- Bên cạnh đó, chúng ta còn có thể áp dụng chuẩn Lp lên biến đổi wavelet của chuỗi thời gian.

Khoảng cách biến đổi thời gian động (Dynamic Time Warping - DTW)

- DTW kéo dãn chuỗi thời gian dọc theo chiều thời gian một cách linh động tại mỗi vị trí.
- Như trong hình sau thì đoạn A, B, C có hình dạng giống nhau, nhưng mỗi đoạn cần được kéo dãn để so khớp với nhau.



(a) Original series (b) Warped series

Figure 3.8: Illustration of dynamic time warping by repeating elements

Source: Data Mining, Charu C. Aggarwal

Các phương pháp dạng ô cửa

- Như trong ảnh sau thì thông tin bị thiếu tạo một khoảng trống khi so khớp.
- Từ đó các phương pháp dạng ô cửa sẽ tách các chuỗi thành các "ô cửa" và dán lại độ đo sự tương đồng.
- Khi đó, nếu 2 chuỗi thời gian có rất nhiều mảng giống nhau thì sẽ được xem là giống nhau.

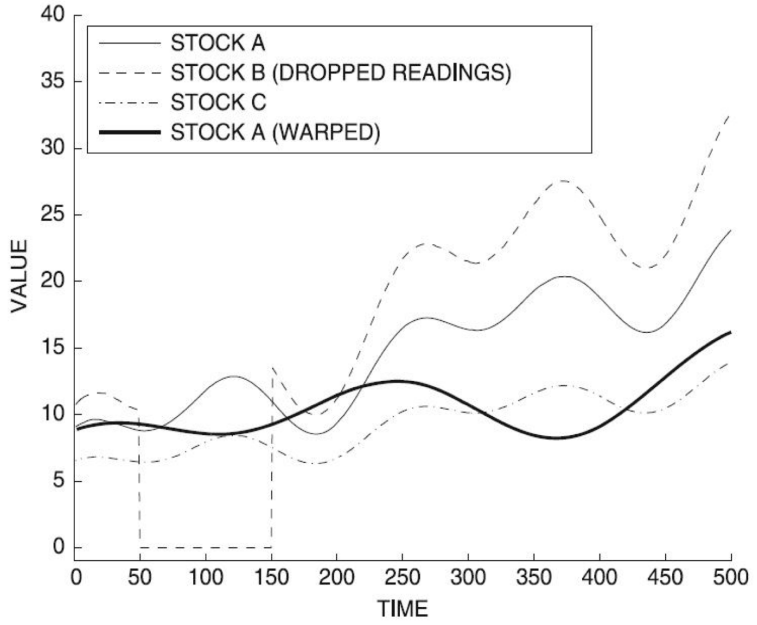


Figure 3.7: Impact of scaling, translation, and noise

Source: Data Mining, Charu C. Aggarwal

Độ đo tương đồng dãy rời rạc

- Các độ đo tương đồng dãy rời rạc dựa trên cùng nguyên lý với các độ đo chuỗi thời gian.
- Khi có ánh xạ song ánh giữa 2 dãy rời rạc tại các vị trí, nhiều độ đo khoảng cách định tính cho dữ liệu đa chiều có thể được áp dụng.
- Tuy nhiên, trong các ứng dụng thực tế với dãy rời rạc thì các song ánh như vậy thường không tồn tại.

Khoảng cách thay đổi

- Khoảng cách thay đổi giữa 2 dãy là lượng việc tối thiểu cần thiết để biến đổi dãy này thành dãy kia qua các phép biến đổi.

- Khoảng cách này cũng thường được gọi là khoảng cách Levenshtein.

Độ đo tương đồng dãy rời rạc

Dãy con chung dài nhất

(longest common subsequence - LCSS)

- Một dãy con của một dãy là một tập hợp các phần tử lấy từ dãy gốc với cùng thứ tự.
- LCSS là một hàm tương đồng với giá trị cao hơn thể hiện sự tương đồng lớn hơn.

Độ đo tương đồng đồ thị

- Với các đồ thị thì sự tương đồng có thể được đo với nhiều cách khác nhau, tùy thuộc được đo giữa 2 đồ thị hay 2 node trong 1 đồ thị.

- Các đồ thị ở đây được giả sử là vô hướng.

Sự tương đồng giữa hai node trong một đồ thị đơn

- Trong một số lĩnh vực, như mạng thư tịch (bibliographic networks) thì các cạnh được gắn trọng số (weights) và hàm tương đồng được dùng.

- Trong một số lĩnh vực khác, như mạng giao thông thì các cạnh được gắn chi phí (costs) và hàm khoảng cách được dùng.

- Thông thường thì việc chuyển đổi giữa hàm tương đồng và khoảng cách có thể thực hiện bằng các hàm kernel.

Sự tương đồng giữa hai node trong một đồ thị đơn

- Chúng ta có 2 tiêu chí về 2 node tương đồng nhau trong một đồ thị
  - Hai node được nối với đường ngắn. Với tiêu chí này, chúng ta có thể sử dụng các thuật toán tìm đường ngắn nhất.
  - Hai node được nối với nhiều đường. Tiêu chí này gắn liền với khái niệm về tính liên thông (connectivity) trong đồ thị.

Sự tương đồng giữa hai đồ thị

- Một trong các yếu tố khiến việc xác định sự tương đồng giữa hai đồ thì phức tạp hơn là các nhãn của node có thể lặp lại.

- VD: các cấu trúc hóa học với mỗi nguyên tố được biểu diễn với một nhãn mà trong mỗi cấu trúc thì cùng một nguyên tố có thể xuất hiện nhiều lần.

Sự tương đồng giữa hai đồ thị

- Với bài toán này, chúng ta có các phương pháp dựa vào khoảng cách thay đổi đồ thị hoặc dựa vào sự tương đồng cấu trúc con.
  - Khoảng cách đồ thị con chung lớn nhất.
  - Sự tương đồng cấu trúc con.
  - Khoảng các thay đổi đồ thị.
  - Kernel đồ thị.

Hàm tương đồng có giám sát

- Trong thực tế, việc chọn hàm khoảng cách có thể phụ thuộc rất nhiều vào lĩnh vực ứng dụng.
- Một số trường hợp như vậy thì hàm khoảng cách không thể được xác định mà không dựa vào mục đích của người dùng.
- Cách để giải quyết vấn đề này là dựa vào các thông tin xác định về sự tương đồng giữa các đối tượng được cung cấp bởi người dùng.
- Từ đây, chúng ta có thể sử dụng nhiều phương pháp có giám sát.
- Bài toán tìm hàm khoảng cách này cũng có thể được viết dưới dạng bài toán phân loại.